# Sensitivity Study of the Skill of the CPC Week-2 Reforecast Tool to Reforecast Sampling

Melissa Ou, Mike Charles, Dan Collins, Emily Riddle
**CPC**

# Outline

- Goals
- Motivation
- Background
- Methodology
- Results
- Conclusion

# Goals

- Determine the impact of changing the sampling of reforecasts on the skill of real-time week-2 (days 8 - 14) calibrated temperature and precipitation forecasts.
- Evaluate the skill of different reforecast sampling cases using various skill scores.
- Find optimal reforecast sampling case(s) that maximized forecast scores.

# Motivation

- EMC requested verification scores and recommendations for CPC's week-2 reforecast tool for reforecast production at NCEP.
- Minimal reforecast sampling requires less resources for producing reforecasts (EMC) and stats calculation and calibration (CPC).
- NCEP GEFS is expected to be upgraded in early 2014.

# Background

- Previously, ESRL had been producing reforecasts but will no longer be doing this. Goal is to have NCEP continue producing reforecasts.
- CPC has been using ESRL's reforecast calibrated tool for 6-10 day and 8-14 day from 2003 to present.
- CPC recently created similar reforecast calibration software to use with upgraded GEFS input datasets.
- Literature has shown that using too many reforecasts may cause overfitting of data (Hamill, 2004) so desirable to find optimal sampling configuration.

# Data

New CPC Week-2 calibrated reforecast tool

- Uses real-time GEFS from Feb 15 2012 to present (physics operational during 2012).
- Forecasts statistically calibrated using GEFS reforecasts, ensemble linear regression method (Unger, 2009).
- Probabilistic tercile categories of temp and precip (T&P)
- Total available reforecasts from 1985-2010.

Verification scores

- 16 months of week-2 calibrated forecasts. Feb 26, 2012 - June 11, 2013
- Mean scores over time, spatially aggregated over the CONUS
- Station observations (205 temperature stations, 190 precipitation stations)

# Steps

**1**

Generate statistics

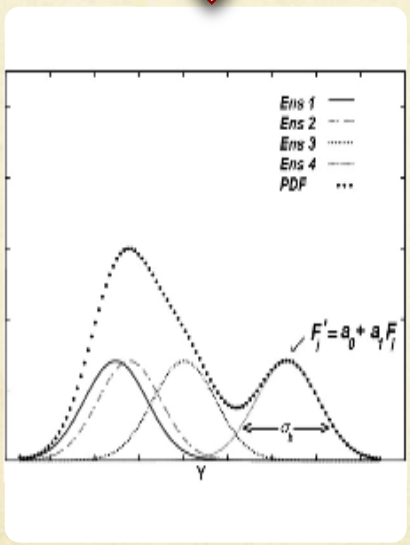
$$\bar{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$$

$$\sigma(x, y) = \mathrm{E}\left[(x - \mathrm{E}[x])(y - \mathrm{E}[y])\right]$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

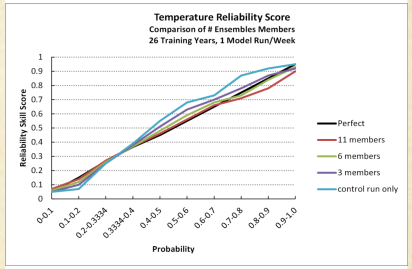
**2**

Calibrate realtime ensemble




**3**

Calculate verification scores




Repeat process for each case

# Methodology
## Sensitivity Test Design

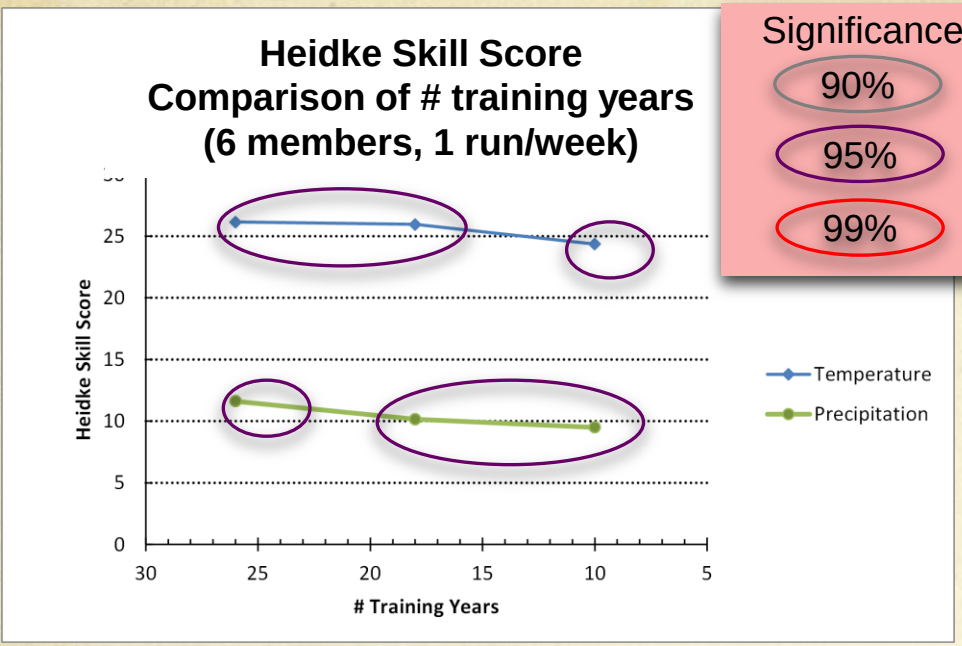
- 11 Cases
- Test sampling of 3 parameters:
  - # training years:
    - 1985-2010 (26 years)
    - 1993-2010 (18 years)
    - 2001-2010 (10 years)
  - # ensemble members:
    - 11, 6, 3, 1 member(s)
  - Model run frequency (times per week)
    - Daily, twice, once a week

# Methodology

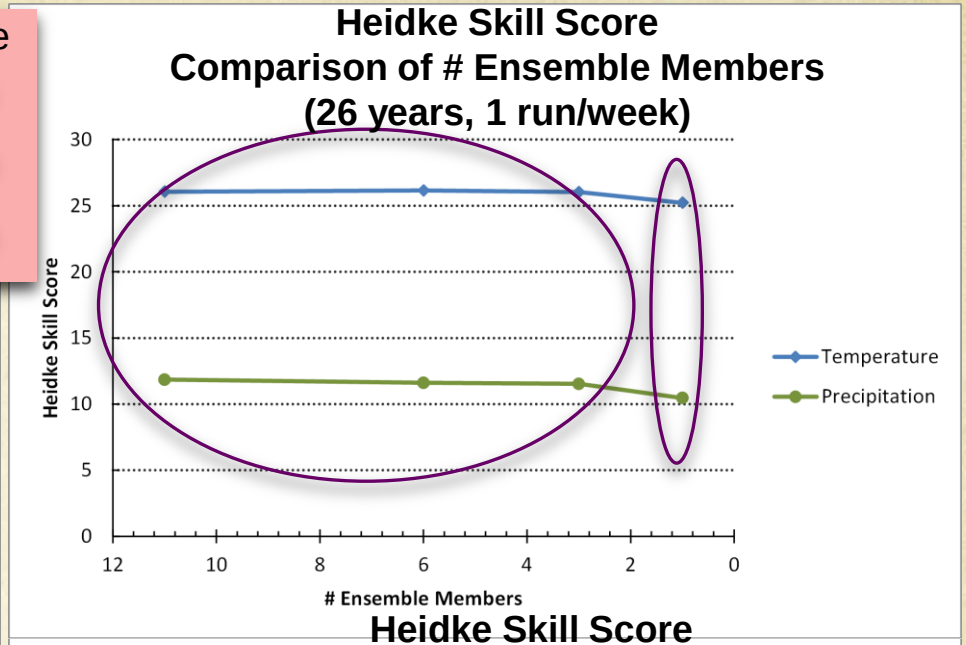
## Evaluating sensitivity of skill

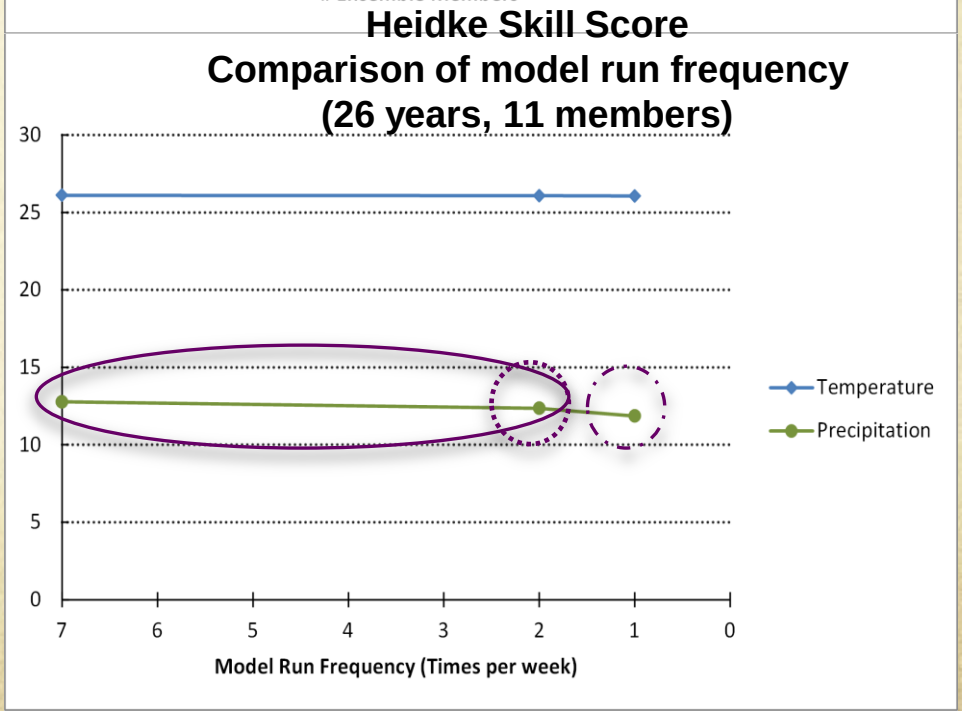
- Skill evaluated using 3 different types of skill scores for verification
  - Heidke, Rank Probability Skill Score (RPSS), and reliability skill scores.
- Created line plots, histograms, and reliability diagrams
- 1-tail two-sample t-test for correlated data to determine significance of mean skill differences (over 16 months of score data).

# Results - Heidke Skill Scores




Significance
90%
95%
99%

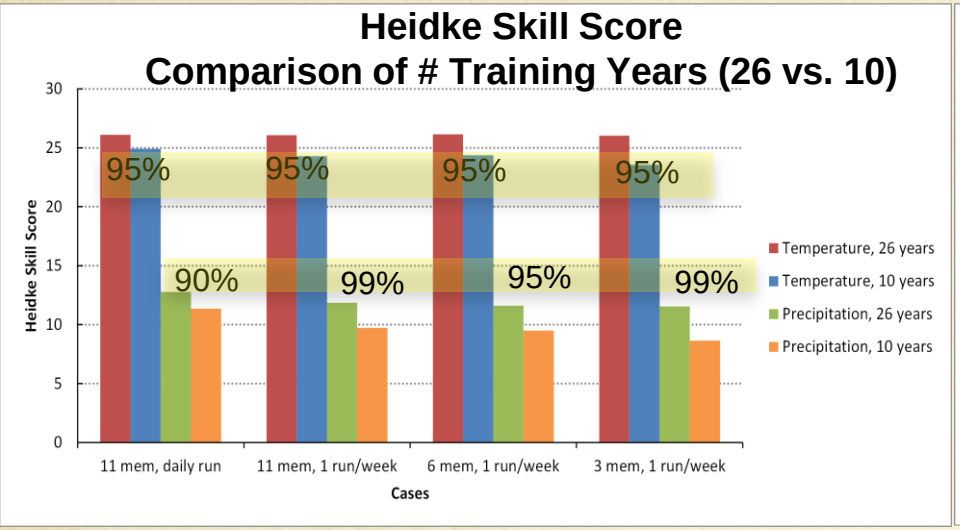


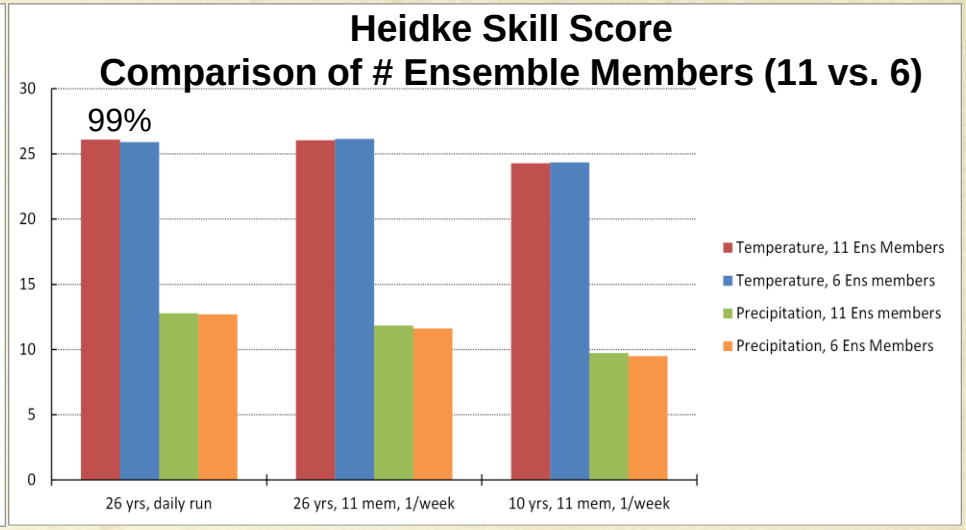




- # Training years shows greatest impact.
- Some differences regarding impact on skill between T&P.
- Precip is more sensitive than temp to training years and model run frequency.

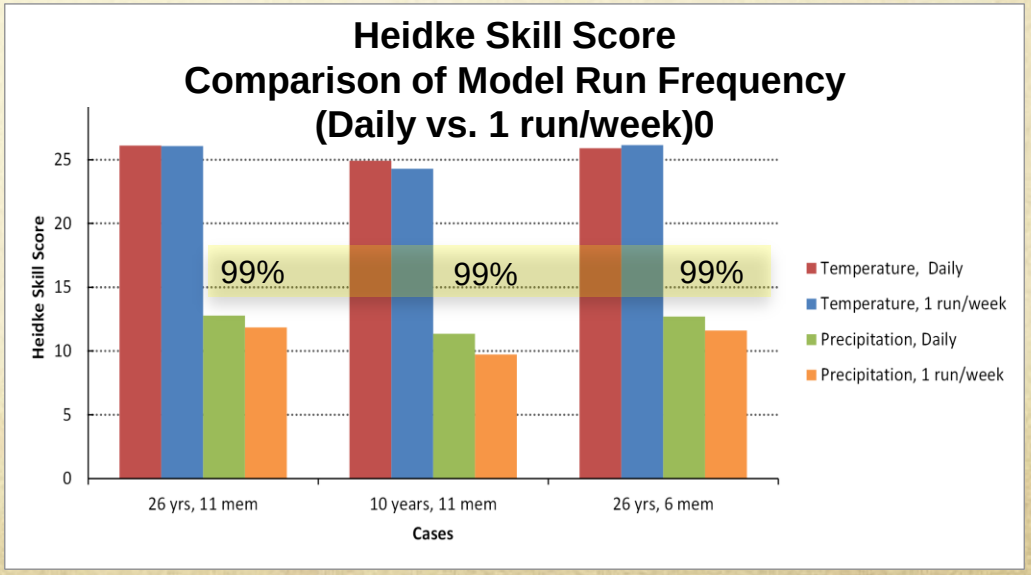
# Results - Heidke Skill Scores

Significance level printed above compared values (>=90%)

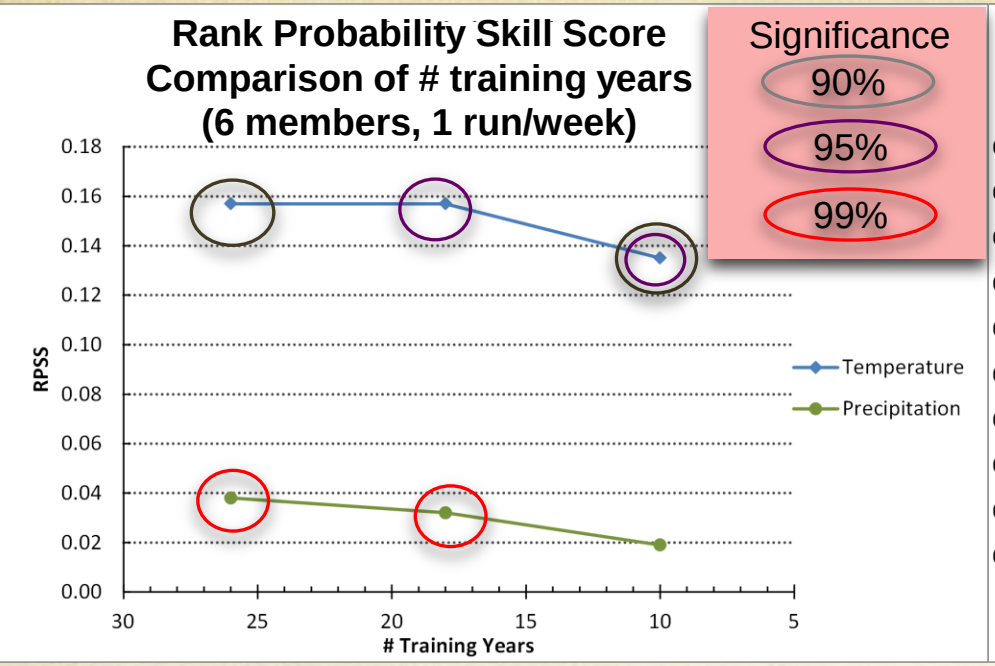


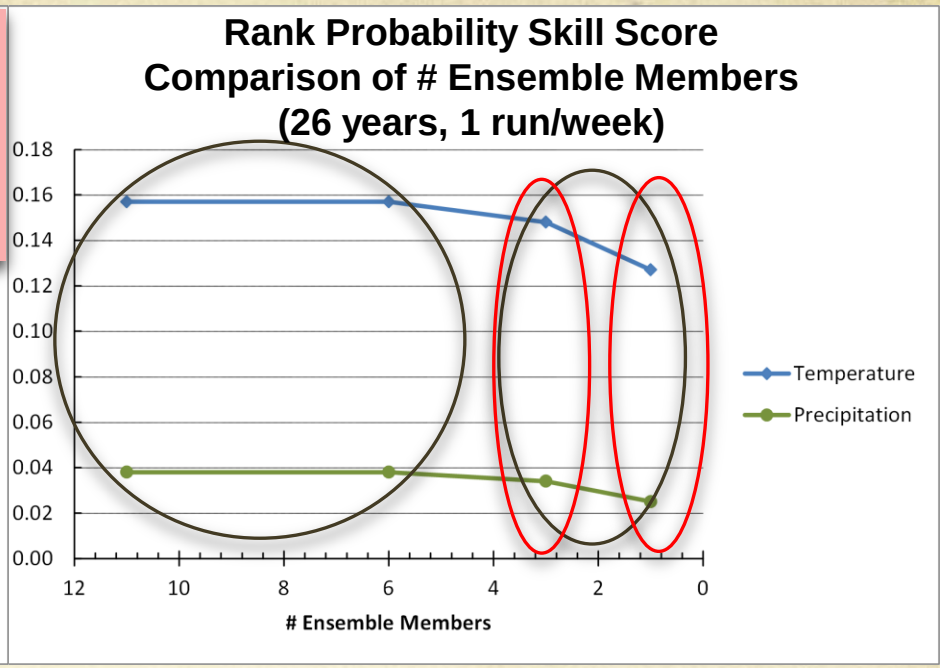




- Training years show the greatest difference in skill for both temperature and precipitation
- # Ensemble members showed least impact

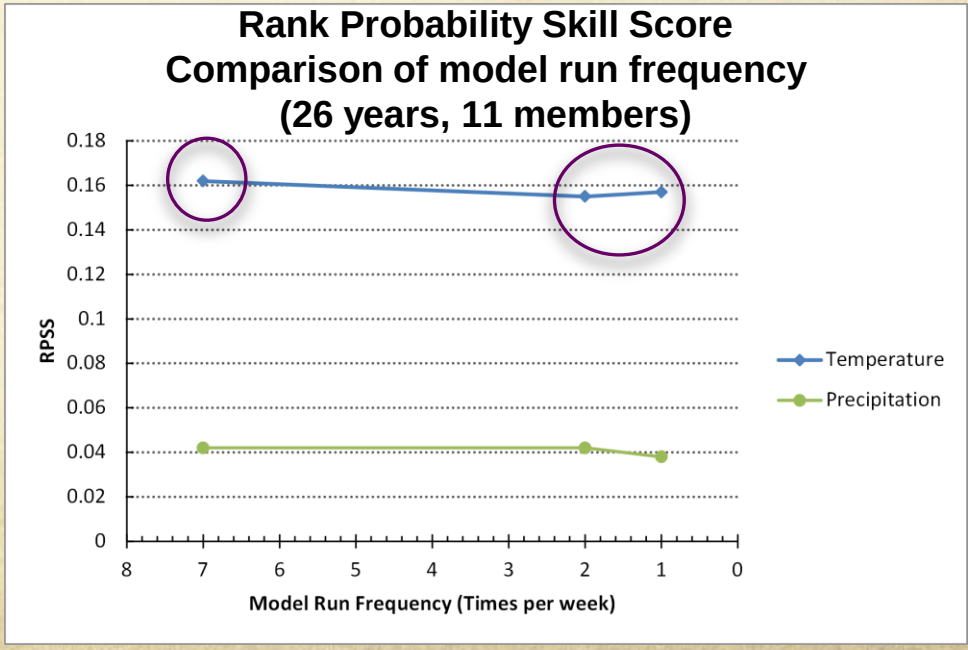


# Results - RPSS



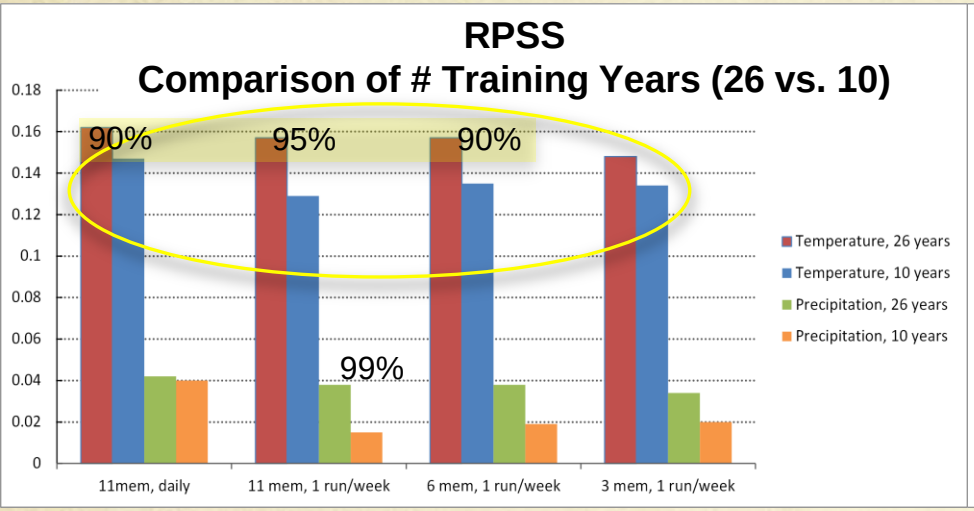



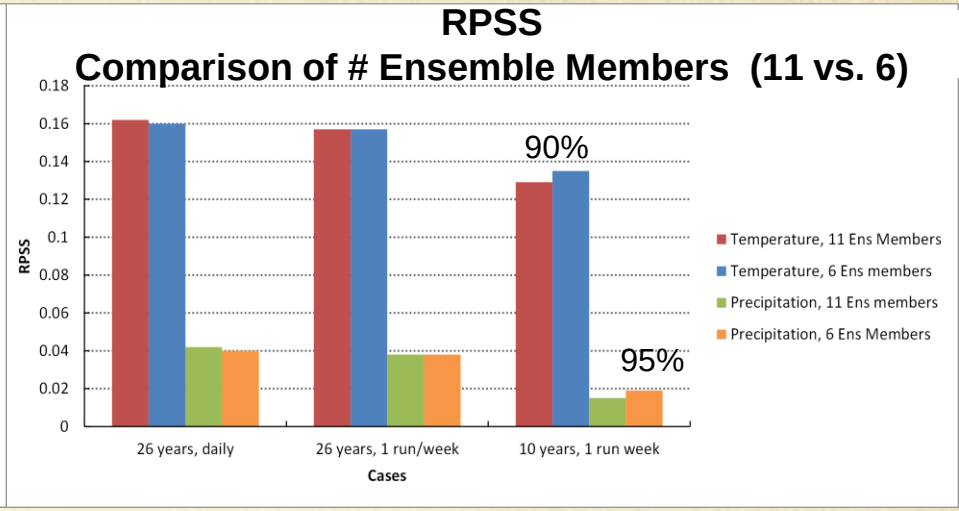

- Steepest drop-off in skill for T&P is from changing # training years
- Least skill impact by changing the model run frequency



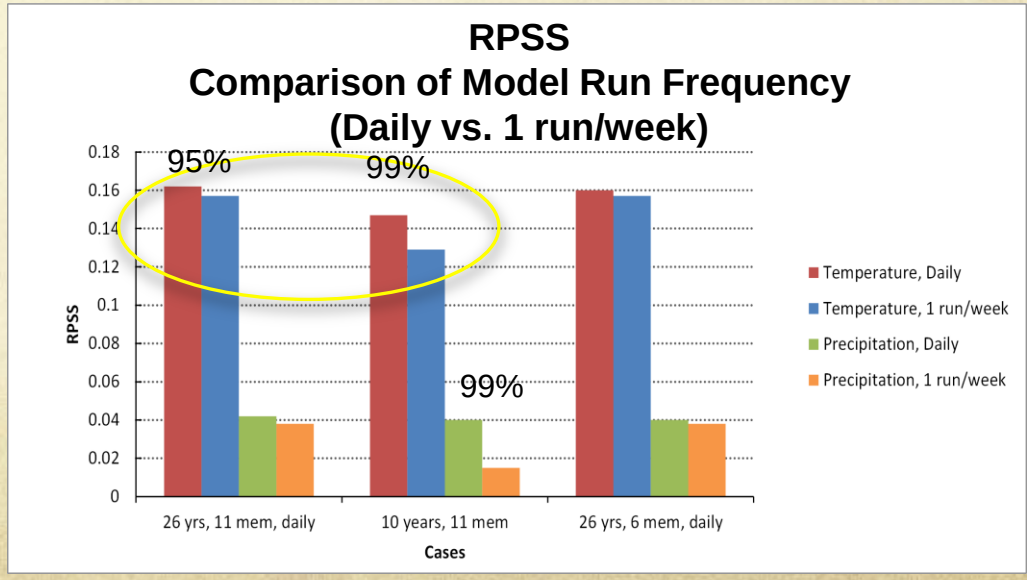
# Results - RPSS

Significance level printed above compared values (>=90%)

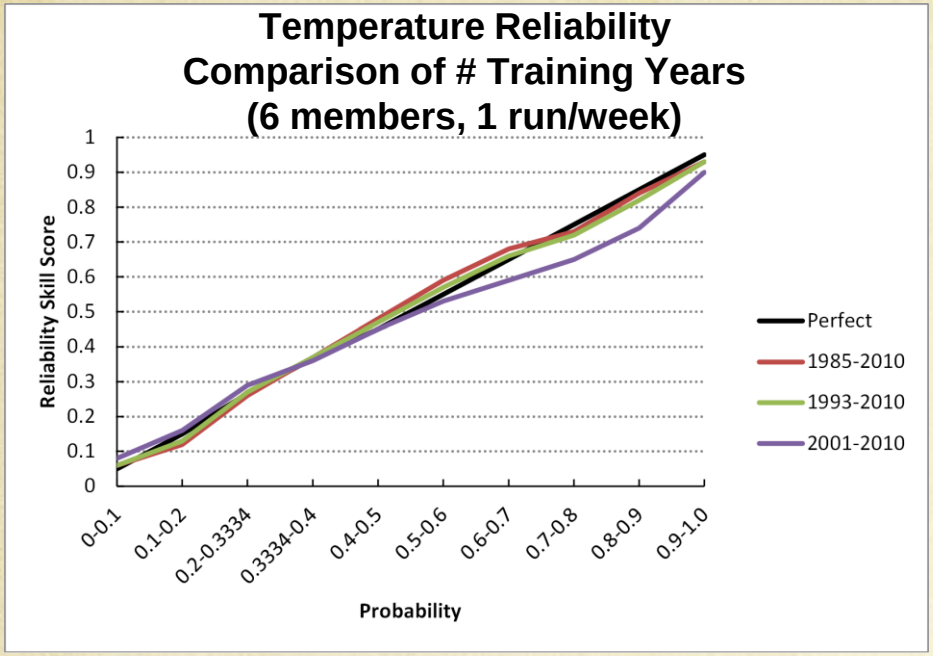


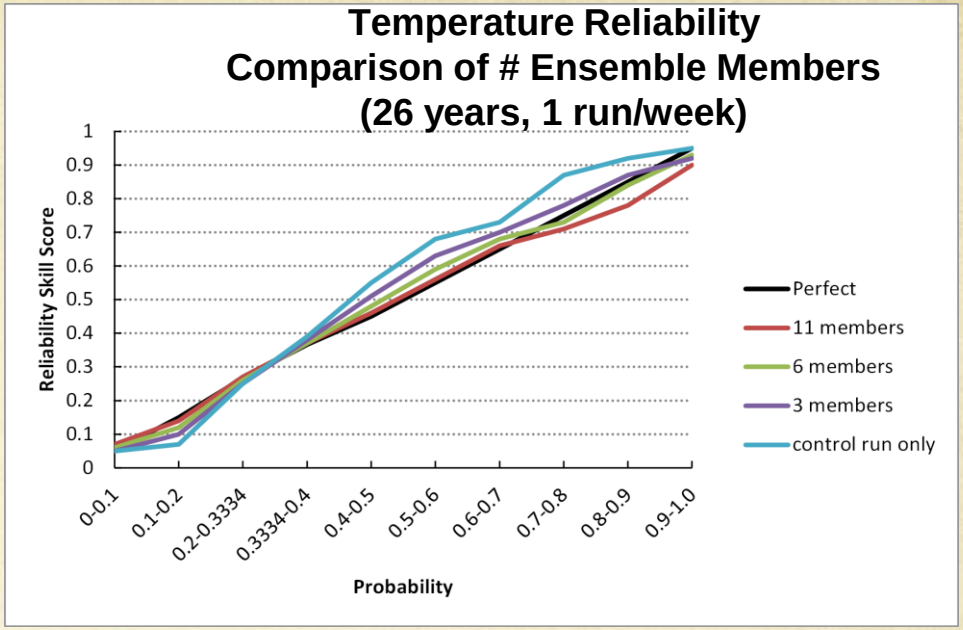




- Similar results as Heidke
- Training years show the greatest difference in skill with significance for T&P
- # Ensemble members showed least impact

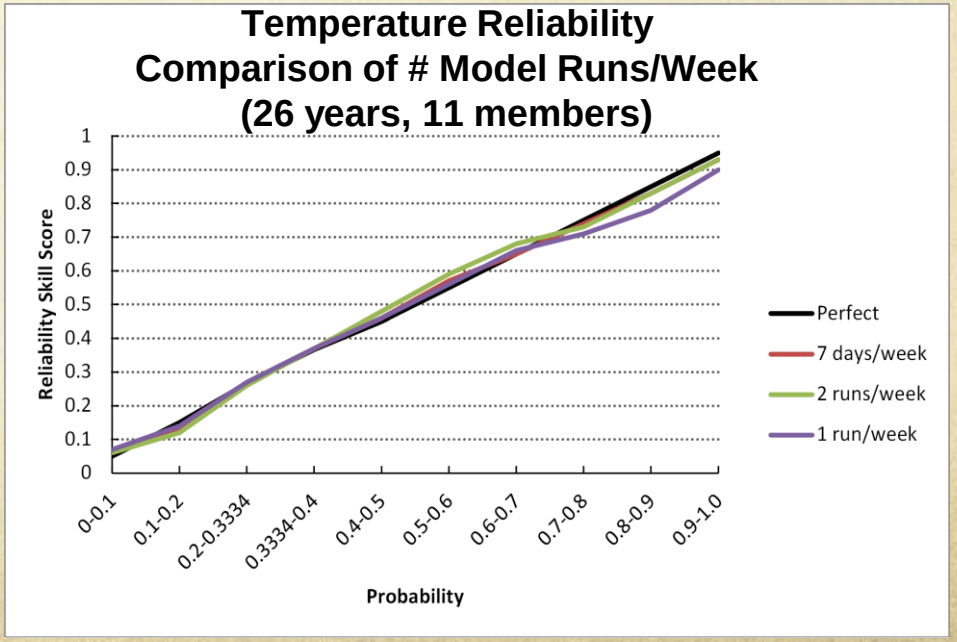


# Results - Temperature Reliability

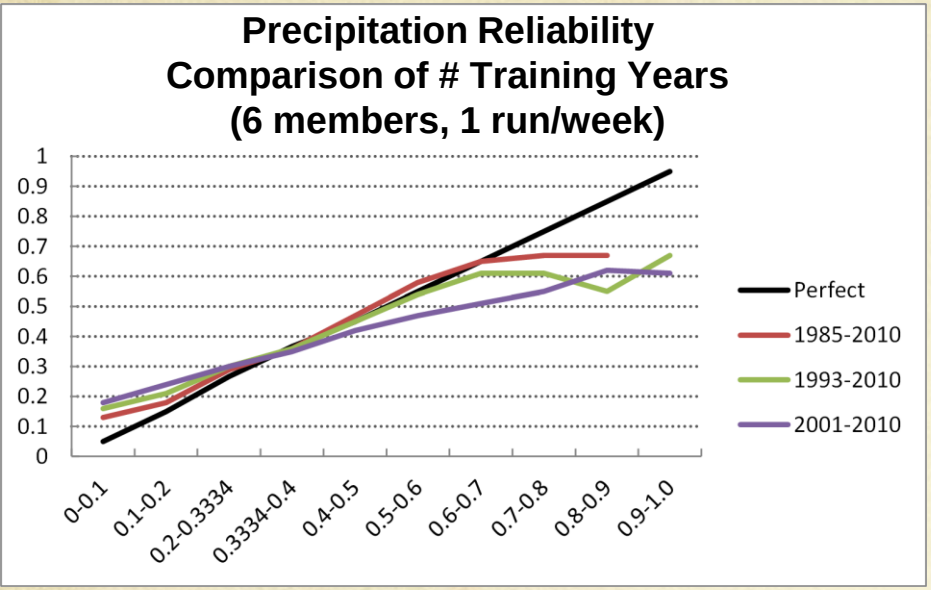


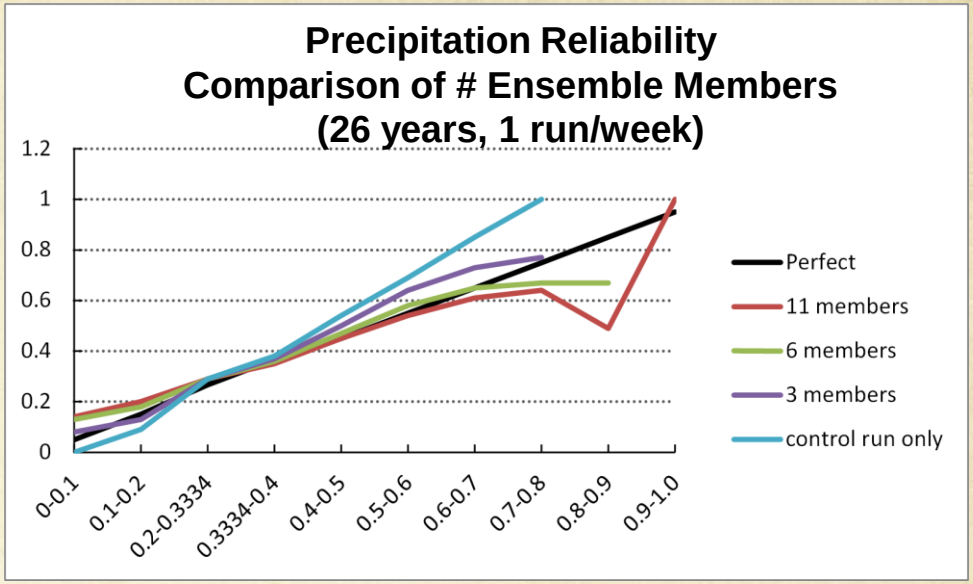




- Reasonable reliability
- Greatest loss in skill from decreasing to 10 training years and only using the control run member.

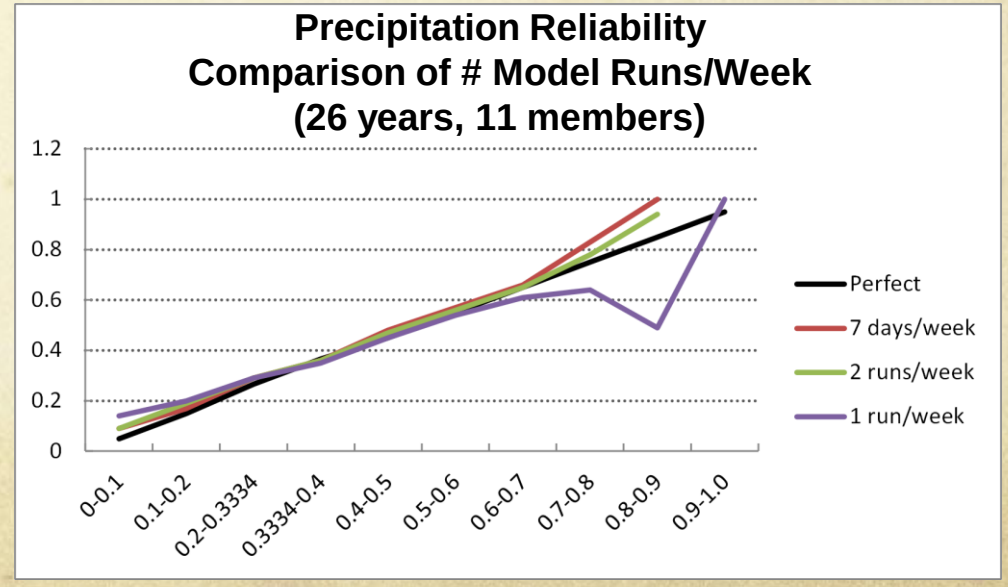


# Results - Precipitation Reliability



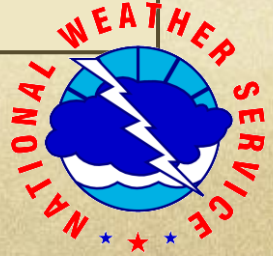

- Reliability skill more sensitive for P than T
- Similar results as temperature



# Conclusions

- Week-2 T&P skill is most sensitive to the number of training years, and least sensitive to model run frequency.
  - T&P both show decreases in Heidke of ~2 and RPSS ~ 0.02 using 10 instead of 26 training years.
- Some differences in the skill of T&P regarding reforecast sampling, but overall similar type of impact.
- It is possible to produce skillful week-2 forecasts using a smaller configuration of model reforecasts
- Using lower configurations of some parameters may improve forecast skill, due to overfitting

Thank You!
Comments? Questions?

# CPC’s report to EMC

- CPC’s report to EMC regarding reforecast production
  - min 20 years, but preferably 30 years, plus recent years (2011-2013) following what we currently have
  - each reforecast should contain at least 5 ensemble members
  - Daily reforecasts provide some benefit and are preferable, but loss of skill using 1 run/week or bi-weekly reforecasts is tolerable given sufficient # training years.
- Our proposed reforecast configuration would cost approximately 26% of the computing of real-time ensemble forecasts.

# Notes and potential future work

- EMC has currently decided to delay production of reforecasts
- We would like to evaluate similar sensitivity studies for the week-2 probabilistic extremes project at CPC using GEFS reforecasts.
- It is possible that the number of training years would be most important in capturing a sufficient number of extreme events
- Currently working with WPC to apply reforecasts for bias-correction forecast tools for QPF, etc. (winter weather desk)
- 6-10 day forecasts would be expected to have similar results (Hamill et al. 2004)

# References


- Hamill, T. M., J. S. Whitaker, and X. Wei, 2004 Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts. *Mon. Wea. Review,* **132**, 1434-1447.

- Unger, David A., Huug van den Dool, Edward O'Lenic, Dan Collins, 2009: Ensemble Regression. *Mon. Wea. Rev.*, **137**, 2365–2379.

# References - skill score equations

**RPSS = 1 - RPS/RPS$_{reference}$ where**

$$<RPS> = \frac{1}{n}\sum_{k=0}^{n}\left[(probB_k - obsB_k)^2 + (probN_k - obsN_k)^2 + (probA_k - obsA_k)^2\right]$$

- Squared error score with respect to the cumulative probabilities for multi-category forecasts and whether or not they even occurred.
- The RPS penalizes forecasts less severely when their probabilities are close to the outcome and more severely when their probabilities are further from the outcome.

***Heidke$_{withEC}$=((numCorrect of nonEC fcsts + numCorrect of EC fcsts) - numExpected)/(count - numExpected)***

where numCorrect of EC fcsts is (num of EC fcsts/numCats) or 1/3 of all EC fcsts when numCats is 3, and numExpected is (count/number of categories) and count is sum of valid EC and non EC fcst-ob pairs. HeidkeWithEC simplifies to HeidkeNoEC * coverage where coverage is (number of non EC fcsts/count).

- The Heidke score utilizes the number of correct and incorrect category hits.

# References - skill score equations

**Reliability**

Reliability for each bin for all categories together is:

reliability = (# obs A / # fcst A) + (# obs B / # fcst B) + (# obs N / # fcst N)

where the # fcst of a category for a probability bin are obtained by counting the number of occurrences when there is a forecast for that category with a probability that falls within the probability bin. The # obs of a category for a probability bin are obtained by counting the number of occurrences where the forecast within that probability bin had that category.

# Reference notes

Hamill, 2004 - results showed that 6-10 day MOS fcsts of sfc temp using only ctrl run was comparable to using 15 members although for precip and week-2 differences were larger. Consistent with notion that benefit of ensemble averaging is a function of ratio of predictable signal (ie ens mean anom) to unpredictable noise (ie ens spread).